Date of Request: November 20, 2017

SCIENTIFIC PEER REVIEW: SOLICITATION REQUEST FORM

Reviewer Information	
Reviewer Name: Ben Jessup	**Title: Ecological Analyst**
Email Address: benjamin.jessup@tetratech.com	**Contact Phone #: 802-229-1059**
Employer: Tetra Tech	**Employer Category: consultant** (federal agency, state agency, academic, professional organization/consultant)

Subject Matter: Biocriteria impairment thresholds

Purpose of Review & Specific Action Required: DEQ is soliciting independent scientific and technical input regarding the biocriteria impairment thresholds that were established in 2012 and are being proposed for 303(d) assessment purposes in the 2018 Integrated Report. Please provide review comments on the questions below.

1. Are Oregon's biocriteria thresholds valid and do they adequately represent the cutoff where aquatic life use is considered to be impaired?
    - If they don't adequately represent the aquatic life use attainment cutoff, what are the limitations of the thresholds and how might they be improved?
2. Oregon currently has two thresholds, one for designated use support (e.g., good biological condition, equivalent to reference) and another for designated use impairment (e.g., poor biological condition, dissimilar from reference). This approach of two thresholds creates a third category of potential concern (uncertain biological condition). DEQ has received input from EPA favoring a single threshold approach, resulting in only two categories of beneficial use support (attaining or impaired). Please provide input on which approach is ultimately more technically defensible in your professional opinion.
3. Are Type I and Type II errors sufficiently balanced by the regional biocriteria thresholds?
    - If not, suggest alternatives for balancing Type I and Type II errors.
4. Are there other methods for determining biological thresholds that DEQ should consider?

Timeline for Review Completion: Reviews should be completed and returned electronically to DEQ by December 29, 2017.

DEQ Point-of-Contact for Reviewer	
DEQ Contact Name: Becky Anthony	**Title: Interim Integrated Report Coordinator, Oregon DEQ**
Email Address: anthony.becky@deq.state.or.us	**Contact Phone #: 541-686-7719**

Specific instructions for providing review comments to DEQ:

Reference documents attached to this request are: (1) Chronology of biocriteria assessment in Oregon (2) Biocriteria methodology summary; and (3) PREDATOR technical report.

Reference and repeat site data used in the PREDATOR model are available upon request.

DEQ staff are available to answer questions, provide additional information or clarifications. Questions should be directed to Becky Anthony (see contact information above).

Please provide peer review comments to DEQ electronically to integratedreport@deq.state.or.us by December 29, 2017.

DEQ follow-up and use of review comments:

DEQ will compile all of the comments received and may reach out to reviewers for explanatory purposes. Comments will be summarized and used to inform revisions to Oregon's biocriteria assessment methodology.

Comments on subject matter reviewed (please attach additional pages as needed)**:**

1. Are Oregon's biocriteria thresholds valid and do they adequately represent the cutoff where aquatic life use is considered to be impaired?

- If they don't adequately represent the aquatic life use attainment cutoff, what are the limitations of the thresholds and how might they be improved?

Yes, the thresholds are valid because they were derived from standard and acceptable analysis methods. Representation of impairment is relative to the quality of the reference sites and the specificity of site classification. Reference sites were presumably the best available – the reference site identification process is still to be reviewed. Site classification was dependent on reference sites available per region – so the specificity might be adequate in areas with low disturbance pressures and might be uncertain in areas with higher general pressure and greater variety of stream settings.

I am an advocate and practitioner of the Biological Condition Gradient for evaluating biological sample integrity. Oregon is participating in a BCG for the Puget Sound and Willamette Valley and will benefit from this process in terms of interpreting biological thresholds in that region and in judging the applicability of the process statewide. The BCG allows for broad expert judgment of the significance of index values in relation to ecological values, so that the thresholds can be crosswalked to narrative and broadly understood levels of biological conditions. Until a biological expert consensus is formulated through a concerted review of a range of samples, the thresholds are valid as relative indicators, but not yet as qualified and interpretable standards for integrity or impairment.

This review process is a great first step towards gaining expert consensus on threshold significance. Building upon this in a BCG calibration statewide would be a valuable progression. Sorry that this might not appear objective, because I am an advocate and practitioner of the BCG. I would be glad to introduce the concepts to this review team if there is interest.

2. Oregon currently has two thresholds, one for designated use support (e.g., good biological condition, equivalent to reference) and another for designated use impairment (e.g., poor biological condition, dissimilar from reference). This approach of two thresholds creates a third category of potential concern (uncertain biological condition). DEQ has received input from EPA favoring a single threshold approach, resulting in only two categories of beneficial use support (attaining or impaired). Please provide input on which approach is ultimately more technically defensible in your professional opinion.

More thresholds are better because the refinement of condition levels allows for different management responses. Above the upper threshold, biological conditions are worthy of protection as high quality resources. Below the lower threshold, restoration activities are warranted, depending on recovery potential. In the middle, where conditions have been labelled as "uncertain", they are actually "certainly mediocre". Certainty should be associated with the precision of the index, which will be associated with any index result, not just those in the middle.

A single impairment threshold might be required for the 303d listing, but multiple thresholds are conducive to refined management responses. If an impairment threshold is definitively placed at only one index value, the other threshold could still be used to trigger other management actions. Degradation from above to below any threshold should trigger an appropriate response to restore the better conditions. The "gold standard" of biological assessment is the TALU/BCG framework advocated by EPA and documented in the critical elements evaluation program. This framework includes single impairment thresholds in the context of multiple other thresholds.

3. Are Type I and Type II errors sufficiently balanced by the regional biocriteria thresholds?

- If not, suggest alternatives for balancing Type I and Type II errors.

This is difficult to assess until the rigor with which reference sites and conditions can be compared to the rigor for stressed site identification. If the rigor and confidence in those designations are equal, then the Type I and II errors should be equal.

4. Are there other methods for determining biological thresholds that DEQ should consider?

Yes, consider expert consensus as in the BCG calibration. A threshold could be set to coincide with an index value representing the difference between BCG level 4 and level 5 (for example, this is not prescriptive). Alternatively, the threshold could be set using percentiles of reference and then interpreted by placing that threshold index value on the crosswalked BCG scale.

The statistical technique of proportional odds modeling is another way to look at probabilities that a certain index value represents a certain assessment category. The models are smooth predictions of index-categories, which is now done using percentiles (which might be subject to capricious index distributions).

These are notes taken while reading the background materials, not organized in response to the questions:

Thresholds of impairment are set according to taxa loss in each regional class. This is based on the percentile of reference in each class. However, the taxa loss in one region (the Marine Western Coastal Forest) is only allowed to be 15% while the loss in the Northern Basin and Range is allowed to be 50% (Category 3C). Based on reference distributions, this suggests that the NBR has a much more variable reference condition than the MWCF. This might be true and might be a model-driven reality of threshold setting, but it also suggests that there are unequal expectation for conditions in the regions, based only on empirical limitations and maybe not on an effort to reduce taxa loss at a minimal level. I will read more – but if 15% taxa loss is unacceptable, why should 50% be acceptable elsewhere? Are the differences explicable because of different ecological mechanisms of taxa loss and replacement? That would give more confidence that equal expectations are set among regions.

Are the additional 5 replicate samples for sites of potential concern taken in one season, or one visit, or are they spread over multiple years? It seems a longer time period would be a better estimate of overall conditions. In practice, multiple replicates were rarely collected. Were 5 replicates decided based on precision analysis or on examples? In Massachusetts, replicate analysis was conducted to arrive at a recommendation that 5 replicates upstream and downstream of discharges would be sufficient to detect a change of x% in metrics. As an example, this is not quite transferable to a reference condition approach, but is similar in the number of replicates.

The 2012 biocriteria modification table appears to be the same as explained for the 2010 biocriteria (except that category 3c and 5 seem to be interchangeable). Am I missing something?

The proposal for 2018 is that there will be no Category 3b (insufficient data) and that everything above the previous impairment threshold will be attaining. Is this the primary reason for this outside review?

"Detrimental changes in resident biological communities are a form of pollution" – Interesting concept – though I don't yet see why this is relevant. I think of impaired biological communities as evidence of other pollutants, except in the case of exotic invasives. The point seems to justify the separation of causation of stressor effects when no other pollutants are identified.

Reference: There is inadequate documentation of reference site identification for a thorough evaluation. Were reference site criteria consistent statewide? Were they based on best available or were criteria more restrictive? These questions qualify responses regarding the adequacy of percentile thresholds.

Date of Request November 20, 2017

SCIENTIFIC PEER REVIEW: SOLICITATION REQUEST FORM

Reviewer Information	
Reviewer Name: Camille Flinders	**Title: Aquatic Biology Program Manager**
Email Address: cflinders@ncasi.org	**Contact Phone #: 360-293-4748 ext. 21**
Employer: National Council for Air and Stream Improvements	**Employer Category: non-profit** (federal agency, state agency, academic, professional organization/consultant)

Subject Matter: Biocriteria impairment thresholds

Purpose of Review & Specific Action Required: DEQ is soliciting independent scientific and technical input regarding the biocriteria impairment thresholds that were established in 2012 and are being proposed for 303(d) assessment purposes in the 2018 Integrated Report. Please provide review comments on the questions below.

1. Are Oregon's biocriteria thresholds valid and do they adequately represent the cutoff where aquatic life use is considered to be impaired?
    - If they don't adequately represent the aquatic life use attainment cutoff, what are the limitations of the thresholds and how might they be improved?
2. Oregon currently has two thresholds, one for designated use support (e.g., good biological condition, equivalent to reference) and another for designated use impairment (e.g., poor biological condition, dissimilar from reference). This approach of two thresholds creates a third category of potential concern (uncertain biological condition). DEQ has received input from EPA favoring a single threshold approach, resulting in only two categories of beneficial use support (attaining or impaired). Please provide input on which approach is ultimately more technically defensible in your professional opinion.
3. Are Type I and Type II errors sufficiently balanced by the regional biocriteria thresholds?
    - If not, suggest alternatives for balancing Type I and Type II errors.
4. Are there other methods for determining biological thresholds that DEQ should consider?

Timeline for Review Completion: Reviews should be completed and returned electronically to DEQ by December 29, 2017.

DEQ Point-of-Contact for Reviewer	
DEQ Contact Name: Becky Anthony	**Title: Interim Integrated Report Coordinator, Oregon DEQ**
Email Address: anthony.becky@deq.state.or.us	**Contact Phone #: 541-686-7719**

Specific instructions for providing review comments to DEQ:

Reference documents attached to this request are: (1) Chronology of biocriteria assessment in Oregon (2) Biocriteria methodology summary; and (3) PREDATOR technical report.

Reference and repeat site data used in the PREDATOR model are available upon request.

DEQ staff are available to answer questions, provide additional information or clarifications. Questions should be directed to Becky Anthony (see contact information above).

Please provide peer review comments to DEQ electronically to integratedreport@deq.state.or.us by December 29, 2017.

DEQ follow-up and use of review comments:

DEQ will compile all of the comments received and may reach out to reviewers for explanatory purposes. Comments will be summarized and used to inform revisions to Oregon's biocriteria assessment methodology.

Comments on subject matter reviewed (please attach additional pages as needed)**:**

1. Are Oregon's biocriteria thresholds valid and do they adequately represent the cutoff where aquatic life use is considered to be impaired?
    - If they don't adequately represent the aquatic life use attainment cutoff, what are the limitations of the thresholds and how might they be improved?

I have some concerns with the PREDATOR models that prevent me from effectively answering this question. The process by which OR DEQ came to have three regional models is reasonable (i.e. examining model performance at different regional scales). However, the resulting model frameworks have not been validated using a test dataset. In speaking with Shannon Hubler (about this and other aspects of the models), this step was omitted on the guidance of a consulting statistician for reasons we did not discuss in detail, and this information is not presented in the documents provided. Although there is often a reluctance in withholding data in the development stage to maximize sample size, model validation is a crucial step in confirming that the developed models are predictive (using data other than with which they were developed), and in quantifying the degree of uncertainty associated with the models (not the variation of model development data, but uncertainty in predictive capability). This is often done by randomly selecting and withholding 10-20% of the available data, and using these to validate the developed model with remaining data. Repeated many times (i.e. using different subsets of the larger dataset for model development and validation), this process can be used to generate a dataset to statistically evaluate probabilities of true and false predictions. The sample size for the Western Cordillera and Columbia Plateau (WC+WP; n=167) is more than sufficient for this exercise, and although much smaller, can also be performed for the Marine Western Coastal Forest (MWCF). These models may very well be highly predictive, but this is unknown without validation using another dataset.

Additionally, there are acknowledged discrepancies in model precision, which model validation will aid in addressing. Hubler (2008) reports that model precision can be estimated by examine the spread of O/E scores in reference sites as represented by the standard deviation of O/E values, and examining the variation in "O" that is predicted by "E" as represented by the r^2 value in a regression of reference site observed and expected values (page 23/62 in the peer-review document). Standard deviations of ~0.15 reflect acceptable precision for a predictive model, while r^2 values from 0.5 to 0.75 in O/E regression reflects a good model. These two precision evaluation methods are contradictory for the WC+CP model, suggesting that this model may not have the predictive capabilities acceptable for evaluating stream condition. Although precision of the MWCF model is corroborated by the two methods, the distribution of data in Figure 3 suggests that the distribution of residuals would be non-random, and merits further examination (or transparency to stakeholders). When applied broadly, high quality/unimpacted streams are likely to be classified accurately, as are highly degraded streams. However, how streams that are moderately disturbed will be classified (i.e. the grey area) is uncertain. Currently, the magnitude of uncertainty in model predictive abilities is unknown (i.e. the size of the grey area is unknown), and cannot be known without validation.

The limited number of data points for the Northern Basin and Range creates additional challenges. The sample size (n=9) does not lend itself to withholding data for validation purposes, and OR DEQs thresholds may not adequately represent impairment. The purpose of bioassessment is confirm that waterbodies are meeting designated uses, and serve as the basis for future management decisions. Currently, the certainty in this model is insufficient to make assessment and management decisions with a high degree of confidence. Additional data to develop a more robust model that can be validated is necessary to develop a better understanding of the predictive capabilities of the model and associated error.

One final note; the datasets from which these models were developed reflect a single biological collection and do not measure or account for temporal variation that may occur at a site. Although the sampling period is limited to June through October (and functionally shorter depending on the source of flow and stream drying, as evaluated by field samplers; S.

Hubler) within-season temporal variability in macroinvertebrate assemblages can be considerable (see Flinders et al. 2015, and citations within), as can variability within and across study reaches in close spatial proximity in the same stream (e.g. Gebler 2004, Gregg and Stednick 2000, Downes et al 2000). I appreciate that biota at reference sites has been shown to be temporally consistent in some studies. However, in at least one study I am aware of, temporal variability in O/E was high enough to result in variable ecological status assessments across years, and that the use of a single sample may affect model accuracy or lead to erroneous management decisions (Huttunen et al. 2012). The temporal variability in biota against the predictor variables in the MWCF and WC+CP models is unknown, but intra- and inter-annual variation may be relatively high (especially in the context of broad-scale predictor variables that do not change (e.g. longitude) or may not change appreciably except under extreme climate conditions (e.g. mean annual precipitation)). OR DEQ has determined that one sample result is sufficient to evaluate for the assessment using the benchmarks developed from the PREDATOR model (page 7/62), but requires 5 replicate samples to provide sufficient data for status classification (pages 1 and 2/62). There is ample evidence supporting replicate samples for bioassessment purposes, but putting known spatial and temporal variability in the context of macroinvertebrate-environment patterns within the least disturbed sites used in model development is important for developing sound biological benchmarks (e.g. Palmer et al 1997, Mykra et al. 2008).

2. Oregon currently has two thresholds, one for designated use support (e.g., good biological condition, equivalent to reference) and another for designated use impairment (e.g., poor biological condition, dissimilar from reference). This approach of two thresholds creates a third category of potential concern (uncertain biological condition). DEQ has received input from EPA favoring a single threshold approach, resulting in only two categories of beneficial use support (attaining or impaired). Please provide input on which approach is ultimately more technically defensible in your professional opinion.

Oregon's establishment of two thresholds is a reasonable and technically defensible approach for determining designated use impairment, and ultimately recognizes uncertainty in evaluating biological condition. It is interesting that EPA favors a single threshold because the two-threshold approach is supported by EPA guidance documents. These include the Consolidated Assessment and Listing Methodology (2002), which outlines an iterative process for improving states', territories', and authorized tribes' monitoring and assessment programs; and EPA's Guidance on Systematic Planning Using the Data Quality Objectives Process (2006), which provides guidance to develop performance and acceptance criteria (or data quality objectives) that clarify study objectives, define the appropriate type of data, and specify tolerable levels of potential decision errors that will be used as the basis for establishing the quality and quantity of data needed to support decisions. EPA's concern that confusion among stakeholders was caused by a third category where the biological condition was uncertain is valid if this information is presented to stakeholders as described in pages 1-3 of the peer-review documents (I found this section confusing). However, my concern over the thresholds identified by OR DEQ is not related to the monitoring program's ability to support such a framework, but to the lack model validation and quantification of uncertainty to establish the specific thresholds identified (see above).

3. Are Type I and Type II errors sufficiently balanced by the regional biocriteria thresholds?
    - If not, suggest alternatives for balancing Type I and Type II errors.

Although OR DEQ recognizes the need to balance Type I and Type II errors, the basis for selecting the 10^{th} and 25^{th} percentiles as assessment thresholds isn't well documented, and seems arbitrary in the absence of quantification of error rates of the three models (through model validation) or evaluation of within-site temporal variation. An important starting point to establishing acceptable (and transparent) error rates is quantifying the magnitude of uncertainty in the predictability of the models through validation exercises, and to examine the temporal consistency of macroinvertebrates at a subset of reference sites (see above). Without knowledge of these components, it is not feasible to determine if the balance between Type I and Type II errors is appropriate.

4. Are there other methods for determining biological thresholds that DEQ should consider?

Another method for determining biological thresholds that DEQ may want to consider is receiver operating characteristics (ROCs). This analysis has been applied extensively for threshold-based classification problems in fields such as medicine and

meteorology, and is being increasingly used in ecological assessments. The approach uses a standard set of calculations to derive several quantitative measures of the performance of a classification model involving a threshold that divides measured and predicted data into two groups (one having (or predicted to have) an undesired condition and one without the condition), and provides a means of assessing the nature and extent of agreement between the true or measured condition and the model-predicted condition. I am not an expert on this technique, but I include papers authored by my colleague Dr. Doug McLaughlin for your review. Should OR DEQ wish to explore this approach further, Doug may be able to provide guidance and insight in doing so.

References

Downes BJ, Hindell JS, Bond NR (2000) What is a site? Variation in lotic macroinvertebrate density and diversity in a spatially replicated experiment. Aust J Ecol 25:128–139

EPA. 2002. Consolidated Assessment and Listing Methodology: Toward a Compendium of Best Practices. First Edition

EPA. 2006. Guidance on Systematic Planning Using the Data Quality Objectives Process. EPA/240/B-06/001. 120p

Flinders CA, McLaughlin DB, Ragsdale RL. 2015. Quantifying variability in four US streams using a long-term dataset: patterns in biotic endpoints. Environmental Management 56: 447-456.

Gebler JB (2004) Mesoscale spatial variability of selected aquatic invertebrate community metrics from a minimally impaired stream site. J N Am Benthol Soc 23:616–633

Gregg DC, Stednick JD (2000) Variability in measures of macroinvertebrate community structure by stream reach and stream class. J Am Water Resour Assoc 36:95–103

Huttunen K-L, Mykrä H, Muotka T. 2012. Temporal variability in taxonomic completeness of stream macroinvertebrate assemblages. Freshwater Science 31: 423 - 441

McLaughlin DB. 2012. Assessing the Predictive Performance of Risk-Based Water Quality Criteria Using Decision Error Estimates from Receiver Operating Characteristics (ROC) Analysis Integrated Environmental Assessment and Management 8: 674-684

McLaughlin DB. 2015. Assessing the Fit of Biotic Ligand Model Validation Data in a Risk Management Decision Context. Integrated Environmental Assessment and Management 11: 610–617

Mykrä H, Heino J, and Muotka T. 2008. Concordance of stream macroinvertebrate assemblage classifications: how general are patterns from single-year surveys? Biological Conservation 141:1218–1223

Palmer MA, Hakenkamp CC, Nelson-Baker K (1997) Ecological heterogeneity in streams: why variance matters. J N Am Benthol Soc 16:189–202

Date of Request November 20, 2017

SCIENTIFIC PEER REVIEW: SOLICITATION REQUEST FORM

Reviewer Information	
Reviewer Name: Dr. Charles Hawkins	**Title: Director, Western Center for Monitoring and Assessment of Freshwater Ecosystems**
Email Address: chuck.hawkins@usu.edu	**Contact Phone #: 435-797-2280**
Employer: Utah State University	**Employer Category: Academic** (federal agency, state agency, academic, professional organization/consultant)

Subject Matter: Biocriteria impairment thresholds

Purpose of Review & Specific Action Required: DEQ is soliciting independent scientific and technical input regarding the biocriteria impairment thresholds that were established in 2012 and are being proposed for 303(d) assessment purposes in the 2018 Integrated Report. Please provide review comments on the questions below.

1. Are Oregon's biocriteria thresholds valid and do they adequately represent the cutoff where aquatic life use is considered to be impaired?
    - If they don't adequately represent the aquatic life use attainment cutoff, what are the limitations of the thresholds and how might they be improved?
2. Oregon currently has two thresholds, one for designated use support (e.g., good biological condition, equivalent to reference) and another for designated use impairment (e.g., poor biological condition, dissimilar from reference). This approach of two thresholds creates a third category of potential concern (uncertain biological condition). DEQ has received input from EPA favoring a single threshold approach, resulting in only two categories of beneficial use support (attaining or impaired). Please provide input on which approach is ultimately more technically defensible in your professional opinion.
3. Are Type I and Type II errors sufficiently balanced by the regional biocriteria thresholds?
    - If not, suggest alternatives for balancing Type I and Type II errors.
4. Are there other methods for determining biological thresholds that DEQ should consider?

Timeline for Review Completion: Reviews should be completed and returned electronically to DEQ by December 29, 2017.

DEQ Point-of-Contact for Reviewer	
DEQ Contact Name: Becky Anthony	**Title: Interim Integrated Report Coordinator, Oregon DEQ**
Email Address: anthony.becky@deq.state.or.us	**Contact Phone #: 541-686-7719**

Specific instructions for providing review comments to DEQ:

Reference documents attached to this request are: (1) Chronology of biocriteria assessment in Oregon (2) Biocriteria methodology summary; and (3) PREDATOR technical report.

Reference and repeat site data used in the PREDATOR model are available upon request.

DEQ staff are available to answer questions, provide additional information or clarifications. Questions should be directed to Becky Anthony (see contact information above).

Please provide peer review comments to DEQ electronically to integratedreport@deq.state.or.us by December 29, 2017.

DEQ follow-up and use of review comments:

DEQ will compile all of the comments received and may reach out to reviewers for explanatory purposes. Comments will be summarized and used to inform revisions to Oregon's biocriteria assessment methodology.

Comments on subject matter reviewed (please attach additional pages as needed)**:**

1. **Are Oregon's biocriteria thresholds valid and do they adequately represent the cutoff where aquatic life use is considered to be impaired?**
    - **If they don't adequately represent the aquatic life use attainment cutoff, what are the limitations of the thresholds and how might they be improved?**

The thresholds are based on approaches generally similar to those used by many other state water quality agencies. ORDEQ currently uses a reference condition approach in combination with an index of taxonomic completeness (observed to expected ratio – O/E), which measures biological condition as the proportion of taxa expected (E) at specific sites that are actually observed (O). Theoretically, sites in reference condition should have index values of 1, and sites whose index values deviate significantly from 1 are considered to not be in reference condition. Inferences of impairment are based on whether observed O/E values fall below a predetermined threshold value. These thresholds are typically less than 1 and ideally represent an index value below which biological harm (impairment) occurs. However, threshold values are typically based on the uncertainty in estimating index values rather than direct interpretation of the biological significance of index values. Estimating index values with error results in a distribution of reference site values theoretically centered on one with a range of values associated with the magnitude of error associated with the estimates. This error includes both measurement error and the error associated with predicting E. Threshold values are therefore typically set that ideally balance type 1 (false positive) and type 2 (false negative) errors of inference. The specific approach used by ORDEQ is to use two threshold values based on the error structure of the indices: one set at the 10^{th} percentile of reference site values, below which sites are considered to be in non-reference condition (i.e., impaired); and another set at the 25^{th} percentile of reference site scores, above which a site is considered to be fully supporting of aquatic life. Values between the 10^{th} and 25^{th} percentiles are considered to indicate uncertain status.

This approach has both strengths and weaknesses. A potential strength is that the specific thresholds are quantitatively based on an objectively established statistical distribution of biological index values observed across reference sites. A potential weakness is that these statistically determined thresholds may not be informed by direct consideration of their biological significance. Instead, biological interpretations are secondarily derived – e.g., for streams in the Marine Western Coastal Forest region, the 10^{th} percentile of reference sites = 15% taxa loss and the 25^{th} percentile represents 8% taxa loss. ORDEQ did not appear to consider an alternative approach in which thresholds were set based on ecological considerations – e.g., how much taxa loss constitutes unacceptable ecological harm. In my view, decisions regarding thresholds of impairment of aquatic life should be primarily based on ecological reasoning and evidence, and the use of these thresholds should then be subsequently supported by appropriate statistical analyses.

2. **Oregon currently has two thresholds, one for designated use support (e.g., good biological condition, equivalent to reference) and another for designated use impairment (e.g., poor biological condition, dissimilar from reference). The use of these two thresholds creates a third category of potential concern (uncertain biological condition). DEQ has received input from EPA favoring a single threshold approach, resulting in only two categories of beneficial use support (attaining or impaired). Please provide input on which approach is ultimately more technically defensible in your professional opinion.**

In my view, a single threshold approach is difficult to justify on statistical grounds given the uncertainty associated with estimating O/E values (or any other index of biological condition), and I think EPA erred in requesting a single threshold. A single threshold approach will have high rates of both type 1 and type 2 errors, which could be the basis for legitimate challenges to assessments. Moreover, ORDEQ applies no formal statistical analyses in support of drawing inferences regarding whether sites are in either of the 2 (or 3) condition categories. Such tools exist, though. I recommend that ORDEQ staff explore the use of equivalency and interval tests to support their inferences. The following publications should be useful in this regard:

Parkhurst, D.F., 2001. Statistical Significance Tests: Equivalence and Reverse Tests Should Reduce Misinterpretation: Equivalence tests improve the logic of significance testing when demonstrating similarity is important, and reverse tests can help show that failure to reject a null hypothesis does not support that hypothesis. *AIBS Bulletin*, *51*(12), pp.1051-1057.

Kilgour, B.W., Somers, K.M., Barrett, T.J., Munkittrick, K.R. and Francis, A.P., 2017. Testing against "normal" with environmental data. *Integrated Environmental Assessment and Management*, *13*(1), pp.188-197.

3. Are Type I and Type II errors sufficiently balanced by the regional biocriteria thresholds?

- **If not, suggest alternatives for balancing Type I and Type II errors.**

In a qualitative sense, ORDEQ has attempted to balance type 1 and type 2 errors similar to the approaches used by other state agencies and supported by EPA guidance. However, no formal analyses have been applied that identify the specific type 1 and type 2 error rates that ORDEQ achieved in each region. Identifying such quantitative estimates is central to establishing defensible biocriteria and aquatic life use standards.

4. Are there other methods for determining biological thresholds that DEQ should consider?

Use of a conceptual framework (such as the biological condition gradient) in conjunction with input from expert ecologists could help inform ORDEQ regarding what amount of biodiversity loss is still supportive of aquatic life use and what level of loss clearly represents impairment. Once these decisions are made based on biological considerations, the statistical methods mentioned above could be employed to support inferences.

End of comments

Date of Request November 20, 2017

SCIENTIFIC PEER REVIEW: SOLICITATION REQUEST FORM

Reviewer Information	
Reviewer Name: Dr. Ian Waite	**Title: Research Biologist**
Email Address: iwaite@usgs.gov	**Contact Phone #: 503-251-3463**
Employer: USGS Oregon Water Science Center	**Employer Category: federal agency** (federal agency, state agency, academic, professional organization/consultant)

Subject Matter: Biocriteria impairment thresholds

Purpose of Review & Specific Action Required: DEQ is soliciting independent scientific and technical input regarding the biocriteria impairment thresholds that were established in 2012 and are being proposed for 303(d) assessment purposes in the 2018 Integrated Report. Please provide review comments on the questions below.

1. Are Oregon's biocriteria thresholds valid and do they adequately represent the cutoff where aquatic life use is considered to be impaired?
    - If they don't adequately represent the aquatic life use attainment cutoff, what are the limitations of the thresholds and how might they be improved?
2. Oregon currently has two thresholds, one for designated use support (e.g., good biological condition, equivalent to reference) and another for designated use impairment (e.g., poor biological condition, dissimilar from reference). This approach of two thresholds creates a third category of potential concern (uncertain biological condition). DEQ has received input from EPA favoring a single threshold approach, resulting in only two categories of beneficial use support (attaining or impaired). Please provide input on which approach is ultimately more technically defensible in your professional opinion.
3. Are Type I and Type II errors sufficiently balanced by the regional biocriteria thresholds?
    - If not, suggest alternatives for balancing Type I and Type II errors.
4. Are there other methods for determining biological thresholds that DEQ should consider?

Timeline for Review Completion: Reviews should be completed and returned electronically to DEQ by December 29, 2017.

DEQ Point-of-Contact for Reviewer	
DEQ Contact Name: Becky Anthony	**Title: Interim Integrated Report Coordinator, Oregon DEQ**
Email Address: anthony.becky@deq.state.or.us	**Contact Phone #: 541-686-7719**

Specific instructions for providing review comments to DEQ:

Reference documents attached to this request are: (1) Chronology of biocriteria assessment in Oregon (2) Biocriteria methodology summary; and (3) PREDATOR technical report.

Reference and repeat site data used in the PREDATOR model are available upon request.

DEQ staff are available to answer questions, provide additional information or clarifications. Questions should be directed to Becky Anthony (see contact information above).

Please provide peer review comments to DEQ electronically to integratedreport@deq.state.or.us by December 29, 2017.

DEQ follow-up and use of review comments:

DEQ will compile all of the comments received and may reach out to reviewers for explanatory purposes. Comments will be summarized and used to inform revisions to Oregon's biocriteria assessment methodology.

Comments on subject matter reviewed (please attach additional pages as needed)**:**

Comments are provided below each of the original questions provided in bold print.

1. Are Oregon's biocriteria thresholds valid and do they adequately represent the cutoff where aquatic life use is considered to be impaired?

I think the various cutoff or breakpoints in the PREDATOR scores seem reasonable for the MWCF = Marine Western Coastal Forest and WC+CP = Western Cordillera + Columbia Plateau. However, even though I understand the reasoning for lowering the impairment bar to 50% loss of taxa for NBR (Northern Basin and Range), with only 10 Expected taxa and many of them common and/or relatively tolerant, a site could pass even though it only has 1 taxa that is more sensitive or intolerant and rest are the relatively tolerant taxa. Yet, without getting more reference sites for this region, there is not a lot that can be done. I also like the current cutoffs in the ODEQ report that provides a range that is for moderately impaired, for I really don't believe in black and white, attain or impaired thinking. Yes, the moderately disturbed sites should be targeted for further evaluation and maybe put on the list, but they are not the same as the sites that are showing full impairment and likely are the sites that could be the easiest to reverse the impairment through restoration and best management practices in the watershed and therefore probably the first sites that should be selected for such restoration and further evaluation efforts. The sites with the lowest scores are in all likelihood the sites that would take the most amount of effort in restoration and implementation of best management practices to see any change at all, or improvements would require a huge cost and multiple decades to see noticeable changes. Thus, identifying the sites that have just gone below attainment are vitally important.

2. Oregon currently has two thresholds, one for designated use support (e.g., good biological condition, equivalent to reference) and another for designated use impairment (e.g., poor biological condition, dissimilar from reference). This approach of two thresholds creates a third category of potential concern (uncertain biological condition). DEQ has received input from EPA favoring a single threshold approach, resulting in only two categories of beneficial use support (attaining or impaired). Please provide input on which approach is ultimately more technically defensible in your professional opinion.

See discussion about in question 1.

3. Are Type I and Type II errors sufficiently balanced by the regional biocriteria thresholds?

As stated in Question 1, yes I believe they are except for the NBR region, where additional effort should be made to see if other references, possibly even those from other States can be added in to improve the model for this Region.

4. Are there other methods for determining biological thresholds that DEQ should consider?

Just a side point, I think the term thresholds is problematic, breakpoints or cutoff values is more appropriate for this purpose. Threshold expresses an ecological change point that is statistically determined and that is not exactly was is being done with the cutoffs decided upon for the O/E models used here. I do think ideally that multiple samples or multiple years should be evaluated to determine sites that are on the cusp of the established cutoffs and again ideally multiple biotic assemblages (algae, invertebrates and fish) should be evaluated to determine the full extent of impairment and the likely environmental stressors associated with the impairment. Given the above, I do believe that using macroinvertebrates and the PREDATOR scores are appropriate for determining quantitative cutoffs and biological criteria and the most reasonable given funding limitations.

Date of Request: November 20, 2017

SCIENTIFIC PEER REVIEW: SOLICITATION REQUEST FORM

Reviewer Information	
Reviewer Name: Dr. Robert Jan Stevenson	**Title: Professor, Michigan State University; Co-Director, Center for Water Sciences**
Email Address: rjstev@msu.edu	**Contact Phone #: 517-432-8083**
Employer: Michigan State University	**Employer Category: Academic** (federal agency, state agency, academic, professional organization/consultant)

Subject Matter: Biocriteria impairment thresholds

Purpose of Review & Specific Action Required: DEQ is soliciting independent scientific and technical input regarding the biocriteria impairment thresholds that were established in 2012 and are being proposed for 303(d) assessment purposes in the 2018 Integrated Report. Please provide review comments on the questions below.

1. Are Oregon's biocriteria thresholds valid and do they adequately represent the cutoff where aquatic life use is considered to be impaired?
    - If they don't adequately represent the aquatic life use attainment cutoff, what are the limitations of the thresholds and how might they be improved?
2. Oregon currently has two thresholds, one for designated use support (e.g., good biological condition, equivalent to reference) and another for designated use impairment (e.g., poor biological condition, dissimilar from reference). This approach of two thresholds creates a third category of potential concern (uncertain biological condition). DEQ has received input from EPA favoring a single threshold approach, resulting in only two categories of beneficial use support (attaining or impaired). Please provide input on which approach is ultimately more technically defensible in your professional opinion.
3. Are Type I and Type II errors sufficiently balanced by the regional biocriteria thresholds?
    - If not, suggest alternatives for balancing Type I and Type II errors.
4. Are there other methods for determining biological thresholds that DEQ should consider?

Timeline for Review Completion: Reviews should be completed and returned electronically to DEQ by December 29, 2017.

DEQ Point-of-Contact for Reviewer	
DEQ Contact Name: Becky Anthony	**Title: Interim Integrated Report Coordinator, Oregon DEQ**
Email Address: anthony.becky@deq.state.or.us	**Contact Phone #: 541-686-7719**

Specific instructions for providing review comments to DEQ:

Reference documents attached to this request are: (1) Chronology of biocriteria assessment in Oregon (2) Biocriteria methodology summary; and (3) PREDATOR technical report.

Reference and repeat site data used in the PREDATOR model are available upon request.

DEQ staff are available to answer questions, provide additional information or clarifications. Questions should be directed to Becky Anthony (see contact information above).

Please provide peer review comments to DEQ electronically to integratedreport@deq.state.or.us by December 29, 2017.

Date of Request: November 20, 2017

DEQ follow-up and use of review comments:

DEQ will compile all of the comments received and may reach out to reviewers for explanatory purposes. Comments will be summarized and used to inform revisions to Oregon's biocriteria assessment methodology.

Comments on subject matter reviewed (please attach additional pages as needed)**:**

1. Are Oregon's biocriteria thresholds valid and do they adequately represent the cutoff where aquatic life use is considered to be impaired?
    - If they don't adequately represent the aquatic life use attainment cutoff, what are the limitations of the thresholds and how might they be improved?

Oregon's biocriteria thresholds are valid and adequately represent the cutoff where aquatic life use is considered to be impaired.

Oregon's biocriteria thresholds are valid because the frequency distribution approach is a commonly used and accepted approach for setting criteria for environmental conditions related to ecological health. It is based on a determination of the natural variation in expected condition measured as the frequency distribution for reference sites (sites that meet management goals) and establishing a benchmark for unacceptable deviation from expected condition within that range.

Oregon's biocriteria thresholds adequately represent the cutoff where aquatic life use is considered to be impaired for two main reasons. 1) The current thresholds are within range of what other states have used routinely to establish thresholds for non-toxic attributes of ecological systems. When states use percentiles of frequency distributions to develop criteria, the 25th, 10th, and 5th percentiles are the most common percentiles used for attributes positively related to expected condition. The 75th, 90th, and 5th percentiles are used for ecological attributes negatively related to expected condition. Oregon's biocriteria thresholds adequately represent the cutoff where aquatic life use is considered to be impaired for many reasons. 2) The standard deviation in O/E for reference conditions were 0.12 and 0.15. Therefore, sites should have natural variation that was commonly greater than the 7% and 8% taxa-loss criteria for the two ecoregions. Thus, 25th percentile thresholds are protective of a relatively high level of biological condition. The 10th percentiles are less protective, and would allow for substantial degradation in condition before impairment was identified and sites were listed on the 303(d) list.

As with almost anything, thresholds have limitations for protecting aquatic life use. But that does not mean the thresholds are not adequate. That means they could be better, which DEQ seems to recognize with their plans to gather more data to improve models, to test different statistical methods for modeling, and to use multiple biological assemblages for assessments of biological condition (Hubler 2008). Cases where sites do not fit into ecoregions that can be modeled well need to be addressed, as is recommended in Hubler (2008). If Oregon wants better assessment of aquatic life, these efforts for model and bioassessment improvements should be funded.

In addition, questions could be raised about sufficiency of Oregon's biocriteria thresholds for protecting aquatic life use from impairment because it is just based on a frequency distribution without more detailed goals for management to support of other ecosystem services. Is protecting about 90% of the species in a habitat sufficient or too restrictive? Why? These issues become more complicated at intermediate tiers of aquatic life use support, where the best quality of aquatic life use is not supported (e.g. Category 3B, 25% taxa lost) and other ecosystem goods and services could be. More elaborate discussion is beyond the scope of this review.

Commented [JS1]: See second bullet.

2. Oregon currently has two thresholds, one for designated use support (e.g., good biological condition, equivalent to reference) and another for designated use impairment (e.g., poor biological condition, dissimilar from reference). This approach of two thresholds creates a third category of potential concern (uncertain biological condition). DEQ has received input from EPA favoring a single threshold approach, resulting in only two categories of beneficial use support (attaining or impaired). Please provide input on which approach is ultimately more technically defensible in your professional opinion.

There are no technical issues (i.e. sampling, modeling, or statistical) with defensibility of having either one or two thresholds that distinguish designated use support and impairment. The main issue is defining what you are trying to accomplish with assessments. In other words, there are conceptual issues. I wanted to make that distinction clear in case I was misinterpreting the request.

The 2010 policy (with Category 3B included) characterizes biological condition at all sites with an O/E score. The 2010 policy does not characterize all sites as either supporting or not supporting designated use because O/E scores for some sites fall in the intermediate range. These sites falling into the DEQ Category 3B, as was used in the 2010 policy, are either sites that are poorly characterized by a single measure of O/E or they have an intermediate quality of biological condition that indeed falls between what was considered (either implicitly and/or explicitly) impaired or supporting aquatic life use in the 2010 policy. Thus, management strategies for this class of sites could not be defined because sites in this range of O/E scores neither fail nor meet aquatic life use support benchmarks. Additional sampling and information for some sites, with true condition levels in the intermediate range, will not reconcile this issue.

If DEQ goes with a 2-tier/1-threshold system, then going with the current threshold for supporting aquatic life use (25th percentile) is more restrictive than the 2010 policy because many sites are on the borderline of the 25th percentile; and many sites in the intermediate range of O/E scores will be classified as impaired (i.e. less that the 25th percentile rather than the 10th). For DEQ to use two tiers as in the proposed policy and to achieve the same level of protection that they had planned in the 2010 policy, the threshold separating the supporting and impaired conditions should be lower than the 25th percentile of the frequency distribution of O/E scores at reference sites.

In summary, both the 2010 and the new EPA approaches need improvement to meet goals of original levels of protection established with the 2010 three tier approach and the likely EPA goal of having clear management strategies for all sites.

3. Are Type I and Type II errors sufficiently balanced by the regional biocriteria thresholds?
    - If not, suggest alternatives for balancing Type I and Type II errors.

As alluded to in previous sections, the Type I and Type II errors that I am considering are for the null hypothesis that condition does not deviate from expected reference O/E scores of 1.0. DEQ established reasonably protective thresholds with criteria for rejecting the null hypothesis (i.e. identifying impairment as a significant deviation from expected condition) with the 10th percentile of reference condition. This establishes a low probability that a site will be identified as impaired when it actually meets reference condition (Type I error = rejecting the null when it is true). It also establishes a relative low Type II error (failing to reject the null when it is true) by using the 25th percentile as a threshold designating a site as supporting aquatic life use.

The new EPA-proposed single threshold policy will have the same Type II error, but a higher Type I error, i.e. identifying a site as impaired when it is not, based on the intent of the 2010 policy that allowed for gathering more information and assigning some sites as supporting aquatic life use when additional information showed that.

So the new approach proposed by the EPA will by more protective, but also more overprotective of aquatic life use.

Alternative approaches include:

1) Set a lower threshold for distinguishing support and impairment with a single threshold, say the 40th percentile, to balance Type I and II errors more closely aligned with the 2010 DEQ policy.

2) Consider modifying the threshold to balance Type I and II errors, and: a) set a boundary around the threshold for classifying sites as requiring more information to determine whether aquatic life use is supported or impaired; b) use repeated sampling to characterize condition more precisely for borderline condition with a guideline for when enough information is gather that a characterization has to be made; and c) use other biological metrics for invertebrates and other biological assemblages to assess biological condition. I do not recommend using stressor and land use characterizations as supplementary information because then the assessments of human disturbance become based at least in part on measures of human disturbance – and the assessment becomes circular.

4. Are there other methods for determining biological thresholds that DEQ should consider?

Yes. Some are mentioned above.

Another is to use tiered aquatic life uses. DEQ could designate the intermediate O/E range as an acceptable but lower quality of aquatic life use than the higher O/E range (<25th percentile of reference condition).

Use of a combination of the approaches mentioned could help resolve the issue identified by EPA, that decisions about condition of many sites are not determined using the 2010 policy.

Reference:

Hubler, S. 2008. PREDATOR: Development and use of RIVPACS-type macroinvertebrate models to assess the biotic condition of wadeable Oregon streams (November 2005 models). State of Oregon Department of Environmental Quality DEQ08-LAB-0048-TR version 1.1.

Date of Request: November 20, 2017

SCIENTIFIC PEER REVIEW: SOLICITATION REQUEST FORM

Reviewer Information	
Reviewer Name: John Van Sickle	**Title: Consultant, Environmental Statistics**
Email Address: vansicklej@peak.org	**Contact Phone #: 541-752-0283**
Employer: Environmental Statistics	**Employer Category: consultant** (federal agency, state agency, academic, professional organization/consultant)

Subject Matter: Biocriteria impairment thresholds

Purpose of Review & Specific Action Required: DEQ is soliciting independent scientific and technical input regarding the biocriteria impairment thresholds that were established in 2012 and are being proposed for 303(d) assessment purposes in the 2018 Integrated Report. Please provide review comments on the questions below.

1. Are Oregon's biocriteria thresholds valid and do they adequately represent the cutoff where aquatic life use is considered to be impaired?
    - If they don't adequately represent the aquatic life use attainment cutoff, what are the limitations of the thresholds and how might they be improved?
2. Oregon currently has two thresholds, one for designated use support (e.g., good biological condition, equivalent to reference) and another for designated use impairment (e.g., poor biological condition, dissimilar from reference). This approach of two thresholds creates a third category of potential concern (uncertain biological condition). DEQ has received input from EPA favoring a single threshold approach, resulting in only two categories of beneficial use support (attaining or impaired). Please provide input on which approach is ultimately more technically defensible in your professional opinion.
3. Are Type I and Type II errors sufficiently balanced by the regional biocriteria thresholds?
    - If not, suggest alternatives for balancing Type I and Type II errors.
4. Are there other methods for determining biological thresholds that DEQ should consider?

Timeline for Review Completion: Reviews should be completed and returned electronically to DEQ by December 29, 2017.

DEQ Point-of-Contact for Reviewer	
DEQ Contact Name: Becky Anthony	**Title: Interim Integrated Report Coordinator, Oregon DEQ**
Email Address: anthony.becky@deq.state.or.us	**Contact Phone #: 541-686-7719**

Specific instructions for providing review comments to DEQ:

Reference documents attached to this request are: (1) Chronology of biocriteria assessment in Oregon (2) Biocriteria methodology summary; and (3) PREDATOR technical report.

Reference and repeat site data used in the PREDATOR model are available upon request.

DEQ staff are available to answer questions, provide additional information or clarifications. Questions should be directed to Becky Anthony (see contact information above).

Please provide peer review comments to DEQ electronically to integratedreport@deq.state.or.us by December 29, 2017.

Date of Request: November 20, 2017

DEQ follow-up and use of review comments:

DEQ will compile all of the comments received and may reach out to reviewers for explanatory purposes. Comments will be summarized and used to inform revisions to Oregon's biocriteria assessment methodology.

Comments on subject matter reviewed (please attach additional pages as needed)**:**

Question 1: The current (2012) thresholds appear to be valid, and they are based on sensible criteria that have been used elsewhere in similar assessments. At present, there are no "gold standard" methods of assessing biological impairment of Oregon streams that are currently available, independently of the PREDATOR model and its supporting data. Thus, it does not appear to be possible to determine, with high confidence, whether or not the 2012 ORDEQ thresholds **"…** adequately represent the cutoff where aquatic life use is considered to be impaired." This reality, along with the inherent vagueness of the terminology of "aquatic life use" and "..considered to be impaired", precludes any purely technical challenge to the current thresholds.

For these reasons, I have long believed that such thresholds cannot realistically be expected to accurately define "impairment" or "attainment". However, such thresholds can be valuable as approximate benchmarks for evaluating changes over time and space that are due to resource usage and management activities. For example, one might report the increase, over the last 5 years, in the percent of streams that are designated as "attaining". Even if the exact definition of "attainment" (as determined by some O/E threshold) is questionable, such change estimates accurately quantify upward or downward trends in the overall health of Oregon's streams.

Question 2: I strongly support the current use of two thresholds, separating 3 categories (2,3b, and 5), for the "uncertainty" reasons given by ORDEQ. I believe that Category 3b gives a necessary buffer to allow for the uncertainty in the assessment.

Here is just one example of statistical uncertainty in O/E scores, which has not been formally factored into category thresholds: The thresholds of O/E for 2 of the model regions are based on the mean and SD of O/E from each region's reference sites. Using textbook methods, one could easily calculate 95% confidence intervals for the estimated mean, and the estimated SD. These results could be combined to estimate the uncertainty of the O/E scores that correspond to the "true" percentiles (10 or 25) of the reference distribution. Because the true percentile locations are uncertain, assessments using any estimated percentile are also uncertain.

Unfortunately, it is not feasible to quantify all of the numerous sources of assessment uncertainty, ranging from uncontrollable variability in macroinvertebrate samples, to statistical uncertainties in the O/E index, and then propagate them all into confidence bounds for the final O/E scores. Thus, ORDEQ's strategy of using a middle, buffer Category between "attainment" and "impairment" seems to be a sensible and conservative approach. Note that USEPA's NARS assessment reports likewise specify 3 classes of biological condition (Good, Fair, Poor) for freshwater systems. I believe that ORDEQ's rules for assessment decisions from replicate samples also provide a commonsense treatment of uncertainty.

Question 3: The thresholds themselves are defined by the Type 1 error rate that is deemed acceptable (10^{th} %ile of the estimated reference distribution of O/E scores). That is, one would expect about 10% of newly surveyed sites that are actually in reference condition would erroneously be declared as "impaired" by the chosen Category 5 threshold (Type 1 errors). A 10% rate for Type 1 errors seems sensible to me, and that rate has also been used to set biocriteria elsewhere.

To accurately estimate the actual rate of Type 2 errors (falsely declaring an impaired site to be "reference" or "attained"), one would need to apply PREDATOR and the assessment thresholds to a collection of sites that are independently known to be impaired. The percentage of such sites declared to be "reference" would then be a good estimate of the Type 2 rate. Such an independent estimate is not available. Thus, it is not possible to know whether Type 1 and Type 2 rates are being balanced.

However, one might specify an approximate subset of "known" impairment sites, based, for example, on them having watershed land uses or other attributes (e.g., mining) that are known to be strongly associated with stream impairment. If one assumes that nearly all of such sites are truly impaired, then they could serve as the independently-

determined "impairment" sites mentioned above.

Until some estimate of the Type 2 rate can be made, I think the effect of threshold choice on the Type 1 versus Type 2 balance cannot be resolved. Meanwhile, however, it would be useful to determine which of the 2 types would be more costly, economically and politically. If one type of error is significantly more costly than the other, then a distinct *imbalance* in the 2 error rates might be most desirable.

Question 4: The statistical approach to setting thresholds as percentiles of the reference scores has the virtue of being more objective than expert-judgement approaches. Thus, I suggest retaining the statistical approach. Although its uncertainties can be substantial, most of them can be quantified.

It might be possible to attain a more robust assessment, with smaller Type 1 and 2 errors, by adopting an average, or maybe a consensus, from multiple indicators of macroinvertebrate assemblage condition. In addition to O/E, one could consider, for example, employing an MMI, and also EPT richness, as condition indicators. However, statewide MMI's would require additional development.

The BC index, an alternative to O/E, is yet another index to consider (Van Sickle 2008, JNABS 27, 227-235). The BC index measures the compositional dissimilarity between the Observed and Expected assemblages. Unlike O/E with its 50% cutoff for capture probability, the BC index can use all reference taxa without losing any discriminatory power. In addition, it avoids the ambiguous "enrichment" issue in which O/E can be greater than 1.0. BC could easily be added to existing PREDATOR outputs. On the down side, BC values are not as readily interpretable as O/E.

I recommend a recent case study (Rose et al. 2016, PLos ONE 11(1): e0146728. Doi:10.1371). They have interesting comparisons between O/E and BC performance, and also among 3 different strategies for predicting the Expected assemblage. For example, recent software now facilitates individual predictive models for every taxon, and this appears to create more accurate Expected assemblages than the awkward machinery of RIVPACS-type models. Finally, their Table 7 is a thought-provoking correspondence between various model performance measures and different bioassessment applications.

Comments (continued)

Becky Anthony
Interim Integrated Report Coordinator, Oregon DEQ
Anthoiny.becky@deq.state.or.us

Re: Scientific Peer Review of Biocriteria Impairment Thresholds

To whom it concerns –

I was asked to review the Biocriteria Impairment Thresholds proposed by ODEQ on 11/20/2017. ODEQ requested comments on 4 questions. The responses to those are provided along with various random additional comments on some of the attached documents.

COMMENTS ON QUESTIONS

1. Are Oregon's biocriteria thresholds valid and do they adequately represent the cutoff where aquatic life use is considered to be impaired?
    a. If they don't adequately represent the aquatic life use attainment cutoff, what are the limitations of the thresholds and how might they be improved?

This is a difficult question to answer on a technical level. The question of where ALU is considered impaired is more a policy question informed by science than a scientific question informed by policy. In my opinion, this should largely be based on the wording of the state's narrative, since there is no numeric criterion in regulation. The state's wording is:

"Waters of the State must be of sufficient quality to support aquatic species *without detrimental changes in the resident biological communities*" (OAR 340-041-0011).(italics added)

As written, the question can be restated as "does the 10th percentile of reference adequately represent where *detrimental change in the resident biological community* occurs?" The answer to that question is also a policy question, since *detrimental* has no clear ecological or scientific definition. It is unclear to what, specifically, *detrimental* applies. Is the state concerned about detriment to other taxa, to specific processes, or something else? If it being defined as detrimental to aquatic life use, then the criterion is circular. I have no specific problem with that, since this is an ALU criterion and it is my belief that such criteria are, inherently, circular and that is fine, since the state is not testing hypotheses but rather defining what is acceptable loss of diversity and function. But, without knowing for sure, it is hard to decide. As written, therefore, the question is a very difficult one and the state, in my opinion, would have great deference as long as they more clearly define the object for which they are protecting from detriment. If, for example, the state is worried about protecting populations of rare, sensitive taxa, then the 10th percentile of reference is likely to be insufficiently protective. If it is common, tolerant taxa, then perhaps it is being too protective. Defining what the ecological goals are that the state wishes to protect from detriment would make it easier to respond to this specific question, in part.
The second part needed to answer this question, then, is what has been lost ecologically, at 10% of reference. Instead of an ecologically based threshold, the state has defined a statistically based

threshold based on a subjectively defined reference population (there is no objectively definable reference population). If the state's ALU narrative were written as "Waters of the State must be of sufficient quality to support aquatic species comparable to that expected under reference conditions", then it would be easier to evaluate the state's biocriteria thresholds and, indeed, some states have taken such an approach to their ALU narrative to better align it with their approaches. Such approaches may be arguably more tied to CWA ultimate goals of integrity. I even think ODEQ may be interpreting their narrative this way, since the documents they shared include the statement: "The scientific peer review panel will be tasked with determining whether clear and convincing evidence exists to support the biocriteria impairment thresholds, and that the status of non-attainment ***represents a significant departure from reference or expected condition***,"(emphasis added). The state asks whether the 10th percentile represents a departure from reference or expected condition, but your standard is not written this way – it is written to prevent "detriment". If you were to define detriment as change from reference, then the threshold could be reviewed as such and I would have to conclude that for the MWCF and WCCP, 15 and 22% loss are likely departures from reference since they are more than 1 standard deviation from the average reference score, but that 50% loss in the NBR is likely too noisy to be useful as a measure of difference from reference. With regards to the existing ecologically based standard (detriment), however, accepting ecological definitions of integrity such as those espoused by Frey et al., the questions for states using ecologically based narrative language becomes how well the reference population reflects those with sustainable assemblages and how much change in species composition represents a departure from that not expected in the absence of disturbance for a self-sustaining assemblage. In such cases, variability in species composition of reference sites in space (reference) or over time, would be a defensible basis for a technically defensible criterion as long as a demonstration that the reference condition meets the criteria of integrity has also been made. For such applications, the question is whether the 10th percentile represents comparability to reference or not, and that can be more easily evaluated technically.

Since the state has not chosen a reference based narrative, in my opinion, the state could provide a stronger, scientific argument for what has been lost ecologically at 10% of reference to justify that detrimental change has not occurred; again, since the narrative is defined ecologically and not statistically. Other states with more ecologically based ALU narratives, have used the biological condition gradient (BCG) model to help define the ecological changes associated with values along their biological indicators to help make just such ecological arguments. I think ODEQ would benefit from a similar effort, or, at a minimum, should consider discussing why the changes at 10% represent a detrimental change in or to biological communities? Absent that ecological discussion, it is hard to evaluate technically, in my opinion. From a technical perspective, using the 10th percentile of reference sites using common take (the state uses capture probabilities of 0.5 or greater), means that there can be sufficient loss of taxa occurring including likely many sensitive and rare taxa, before an action is taken. If I am reading the PREDATOR report correctly, and there are only 10 taxa considered (average E of 7.6) in the NBR class for expected richness and the 10th percentile of reference is 0.5, then on average 4 of those 7.6 expected taxa would need to be missing from a site before it is considered impaired. In my professional opinion, it is hard to imagine there not being detrimental changes in any ecosystem before the loss of 50% of the common taxa. Even the author of the PREDATOR report acknowledges that the SE Oregon index should be used with great caution. Ten taxa provides a very low signal and O/E models get very unstable with such low expected taxa richness.

A problem is that taxa are not created equally (e.g., some species of filter feeding simuliid blackflies in high densities can alter the composition of fine particulate organic matter; single perlid predators in rocky mountain streams can control prey populations). So, absent an ecological discussion of what is happening at these thresholds ecologically – what has been lost, what functions therefore might be vulnerable – evaluating the defensibility of the ALU threshold is difficult if not impossible. Adding this defensibility using BCG or professional staff interpretation, would strengthen the argument.

2. Oregon currently has two thresholds, one for designated use support (e.g., good biological condition, equivalent to reference) and another for designated use impairment (e.g., poor biological condition, dissimilar from reference). This approach of two thresholds creates a third category of potential concern (uncertain biological condition). DEQ has received input from EPA favoring a single threshold approach, resulting in only two categories of beneficial use support (attaining or impaired). Please provide input on which approach is ultimately more technically defensible in your professional opinion.

Again, and hopefully I am not being difficult, I am not sure there is a technical question here. I think if you address the comments raised in (1) above, then you address this question as well. Your narrative is black and white, but unfortunately, ecological response is not. If one believes the BCG model of how ecological changes occur, then everything we understand about how stressors affect streams is that there are very rarely clear step functions in response. Rather, the technical literature all indicates that most biological responses to stressors in streams are gradual. Therefore, there is no clear technical line of "detriment on this side, not on this side". This distinction is only a policy one. For an ecologically based definition, what can be used, technically, to inform such a decision would be to discuss how much taxa loss can be sustained without a change in function (e.g., litter processing, primary productivity, nutrient uptake) or structure (e.g., the loss of the n^{th} species results in a dramatic shift in species composition with little likelihood of recovery – the composition shifts into a new stable state). The state does not create or replicate this gradual response condition be using two thresholds, they only create the need to justify those two values. At some point, some needs to make the policy decision that "this much change is detrimental" and that should likely be based on a more detailed discussion of what the state desires to protect and how a specific value represents a detriment to that protection. For some, that would 51% likelihood of a detrimental change, for others 99%.

As for the question of whether two thresholds or one is more technically defensible, since your narrative does not speak to three conditions but only two (one side is detriment and the other not), then one threshold would appear to me to be more defensible. Creating a gray or middle zone within which one is unsure of attainment is not more defensible. To say that less than 25% of reference is concern and less than 10% is detriment is, in my opinion, confounding your uncertainty. ODEQ is basically asserting that the 10% if the detriment position and the 25^{th} is a concern level. But your standard is written to one thresholds, detriment, not to two.

3. Are Type I and Type II errors sufficiently balanced by the regional biocriteria thresholds?
    a. If not, suggest alternatives for balancing Type I and Type II errors.

You are not testing hypotheses here using randomized controlled experiments of some controllable treatment, so the concepts of Type I and II error seem misplaced. Since you can neither know,

independently, what attainment or impairment is (you are defining it yourself), then you cannot apply a statistical test to it in my opinion. So asking whether such errors are balanced is not relevant, in my opinion. If you had independent measures of attainment, you could look at decision agreement among criteria, but I am not sure type I and II error are relevant. I think, for this setting, the latte may be valuable. If you had sites that were deemed impaired for aquatic life use based on independent measures (DO, pH, ALU metals criteria, etc.) and you compared those with your biologically based thresholds, it would certainly lend some strength to the argument that these thresholds protect against *detrimental* harm. Presumably, your biota respond to these stressors and if you can demonstrate that you observe bioindex scores below your O/E thresholds above the values of these stressors known to cause *detriment*, you can strengthen your argument.

4. Are there other methods for determining biological thresholds that DEQ should consider?

Well, from my discussions above, I think you can guess that using ecological information embodied by the taxa typically protected (and lost) under the proposed thresholds would be valuable in evaluating thresholds. This is most easily done under a BCG modeling effort, but the state has other options. These include state biologists reviewing what sites deemed impaired by such thresholds embody, ecologically, above and below the proposed thresholds and an argument constructed for why such conditions are/are not *detrimental*.

Another approach, as detailed above, would be to refocus the narrative on reference condition and then making the argument for why the 10^{th} percentile is different from reference. I think the latter path would be easier, technically, but maybe not politically.

ADDITIONAL COMMENTS

Below are some additional notes relevant to the documents sent and observations that arose during the review.

A. The O/E model

I think the state would be justified in recalibrating your O/E model. It has been sufficient time and the state has made an effort to collect additional reference data, I think the model should be updated, as recommended in biocriteria guidance. Improvements in O/E modeling (e.g., use of random forest models vs. all subset discriminant analysis, additional validation methods, etc.) as well as the development of a plethora of additional predictor data through StreamCat make updating the model a good idea. The existing models use a surprisingly few number of predictors – 2 in MWCF and 4 in WCCP. Most O/E models use substantially more predictors.

B. Percentile selection

The state argues in their document that the 10th percentile was selected as the threshold for where "detrimental changes in the resident biological communities" have occurred, but there is little in the way of any justification for this, other than that it is a percentage of reference. The concept of detriment implies some adverse impact on the community. What does that mean ecologically? What is happening to streams in OR at the 10th percentile? What changes have occurred to suggest the impacts are detrimental? How can 15% taxa loss in one stream class be an equivalent level of detriment to 50% loss in another? Are streams so plastic in their resilience to stress or in the redundancy of taxa to sustain functions and structure?

C. Replicate sampling

The state makes an argument that DEQ expects some reference sites will score below the 10th percentile. That is true, about 10% of the original population will (☺). However, will they twice? How likely is it that two samples from one stream both score below the 10th percentile? You have replicate samples and it seems you could answer this. Obversely, you have independent chemical measures of ALU (DO, pH, metals, etc.). How likely is it that a site deemed impaired by exceeding those criteria scores above the O/E thresholds? How does that inform your decision-making with regard to thresholds? To use your example, how high would the "type II risk" be? (Again, I think this is not an appropriate use of Type I and II error – but for sake of argument I will use this language). I encourage you to read Doug McLaughlin's papers on decision agreement related to nutrients for ideas on how to compare these thresholds.

D. Confidence in Reference

ODEQ states they are confident that single score below the 10th percentile is not different simply by chance, but rather a true difference in biological condition exists. Why? There is a 10% chance one of your reference site scores below this and we do not know how known impaired sites score, so we cannot really evaluate such decision agreement – but I think you have the data to do it and should think about doing it, as suggested above. Also, it would help to check repeat reference site samples against this assumption. Since you chose a low percentile and are assuming a space for time substitution, one

should assume that a reference site scoring below the 10th percentile twice would have a very low probability of occurring. You can test that.

E. Edits to Narrative Assessment Protocol

"Comparison with Expected Condition: DEQ supports the use of "reference condition" or "expected condition" as the basis for characterizing use support. It is important to note that this concept of use support embraces considerable variation in the biological community. This variability is ~~acknowledged~~ included in developing the biological thresholds."

F. Questions from PREDATOR report

p.9 An argument is made that it is not advantageous to develop a predictive model if too reference sites are available. But the null model still uses reference sites. So, how are there too few to predict, but enough to set a threshold?

p.10 An E of 7.6 is a painfully low signal upon which to base an index. I am not sure such models are defensible. Did you consider decreasing capture probability to increase signal in these sites? How would the model behave with a lower Pc?

p.15 The NBR is haunting me, as I am sure it does Shann. But looking at Figure 2, one can starkly see how the NBR is so different. The E for NBR is not even in the range for MWCF and only barely for MWCF. It indicates how starkly different that index is. I just think there is a real risk that the ALU expectation for the NBR is far different then for the rest of the state and yet, there is only one narrative ALU criterion, so it does not suggest the NBR should have a different expectation. I think ODEQ recognizes this, because they write: "Performance of the Northern Basin and Range (NBR) null model cannot be assessed in the same way as predictive models. By definition, the mean O/E value for the reference sites used to build the null model is 1.0 (Table 3). Precision can be estimated by looking at the SD of O/E values for reference sites (Table 3). The high SD of O/E values for reference sites suggests low precision. Obviously, ***having only nine reference sites in the NBR limits our confidence in our assessment of biological condition in this region***." Then on p. 20 the state writes: "Until DEQ develops a more accurate model for SEOR, I ***recommend using the SEOR null model with caution in bioassessments***." (Emphases in both cases added). I could not agree more with these statements.

p. 19 Turak et al and Clarke et al. references were missing from literature cited.

Respectfully –

Michael J. Paul